Check for updates

Proposal

Disclosure without the Social Penalty: A Trust-Preserving Standard for AI Transparency

Behavioral Science & Policy
1–5
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23794607261483325
journals.sagepub.com/home/bsx
Sage

Oliver Schilke[1] and Martin Reimann[1]

Abstract

Calls for transparency in AI use are increasing. Yet experimental evidence has shown that people trust an actor less after an AI use disclosure,[1,2] which can trigger legitimacy concerns such that the behavior is perceived as atypical and the discloser is seen as insufficiently committed and inauthentic. This AI disclosure penalty can create a policy trap in which blanket disclosure rules can unintentionally penalize conscientious actors and encourage concealment. This article aims to translate recent evidence on the AI disclosure trust penalty into recommendations for a disclosure standard that preserves transparency without undermining the perceived legitimacy of AI-assisted work. The coauthors offer four recommendations: (a) normalizing approved AI use before establishing mandates, (b) tiering disclosure obligations according to the stakes, (c) pairing disclosure with an ownership and accountability statement, and (d) backing mandates with enforcement.

Keywords

AI disclosure, transparency, trust, legitimacy, accountability

Key Points

- Across a series of experiments, disclosing AI use lowered trust in the discloser; adjusting the disclosure's framing helped only modestly.[1]
- Disclosure can backfire because it triggers legitimacy concerns about lack of typicality, commitment, and authenticity.[2]
- Third-party exposure of undisclosed AI use damages trust even more than voluntary disclosure, making concealment a risky strategy with potentially severe legitimacy and reputational consequences if exposed.[1]
- Policy should shift from whether to disclose to how to disclose while supporting legitimacy.

The Disclosure Dilemma

Consider a manager preparing annual performance reviews using a generative AI tool to draft initial language. Following guidance for AI transparency, she discloses the AI assistance at the end of each review. But this disclosure will likely lower employees' trust in this manager, even if the performance feedback is accurate and she thoughtfully revised the initial language.

That finding is not fictional. In one experiment close to this example, legal analysts evaluated a supervisor after receiving an annual performance review. Trust in the supervisor was lower when the review stated it had been prepared by generative AI ($M = 3.80$ on a 7-point scale) than when no such statement was made ($M = 5.00$).[1] And this effect generalizes across settings. Experimental evidence spanning many workplace tasks and professional roles has shown that revealing AI use can produce a substantial trust penalty arising from an audience's inference that the source relied less on appropriate procedures than expected.[1,2]

Trust is the willingness to accept vulnerability to another party's actions.[3] When trust levels fall, people cooperate less, monitor more, and coordinate more slowly as a result.[4,5] A disclosure-triggered drop in trust therefore has considerable economic costs, as well as equity costs arising from selectively penalizing those who comply with disclosure expectations.

This tension between the ethical benefits of transparency and the social costs of disclosure creates a dilemma for regulators and employers. Transparency may reduce the risk of misleading stakeholders. However, AI disclosure mandates

[1]The University of Arizona, Tucson, AZ, USA

Corresponding Author:
Oliver Schilke, The University of Arizona, Tucson, AZ 85721, USA.
Email: oschilke@arizona.edu

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Table 1. Empirical findings, legitimacy concerns, & policy implications.

Empirical finding	Legitimacy concern	Policy implication
AI disclosure lowers trust across tasks, holding work quality constant.	Disclosure can make AI-assisted work seem inappropriate.	Normalize approved AI use through approved-use lists, routine templates, and leaders visibly demonstrating approved AI practices.
Disclosure heightens concerns about typicality, commitment, and authenticity.	Audiences ask whether AI use is normal, whether a human put in the required effort, and whether the output is genuinely owned.	Tier disclosure obligations according to the stakes such as stating what the AI did, what the human did, and who is accountable. Increase disclosure requirements as stakes rise.
Alternative disclosure framings seem to help only modestly.	Disclosure rhetoric alone does not resolve lingering doubts about appropriate use.	Pair disclosure with an ownership and accountability addendum.
Third-party exposure of undisclosed AI use damages trust more than voluntary disclosure.	Concealment creates reputational risks but only if it gets exposed.	Predictably enforce mandates to raise the perceived likelihood of exposure.

can also produce adverse selection. Conscientious actors who comply risk incurring reputational damage, whereas others conceal AI use and avoid repercussions. An informed approach should both preserve the ethical benefits of transparency and reduce unnecessary trust damage.[6,7]

What the Research Shows

To examine how AI use disclosure affects trust, we used a series of experiments in which participants evaluated otherwise similar work products or decisions with one key difference: whether or not AI assistance was disclosed. Participants considered tasks such as workplace messages, performance-related evaluations, financial judgments, and organizational communications. The core pattern was consistent across these settings: When AI use was disclosed, participants trusted the person or organization significantly less, even though the quality of the work itself was held constant.[1,2]

A within-paper meta-analysis conducted across individual experiments estimated the average disclosure penalty as a standardized mean difference of about 0.8.[1] This effect size indicated that trust in the disclosure condition was roughly eight tenths of a pooled standard deviation lower than in the no-disclosure control, which is a large effect by conventional standards. Because the effect recurred across diverse tasks, samples, and forms of AI use, the penalty did not appear confined to any single setting.

Framing the disclosure in different ways helped less than one might expect. In a job application study, trust was lower after a basic AI disclosure than in the no-disclosure condition. Adding that a human had reviewed and revised the statement did not materially improve trust, and even a clarification that AI was used for proofreading only led to trust levels markedly below the no-disclosure condition. Thus, disclosure wording mattered at the margins, but it did not remove the penalty.[1] The practical implication is that tweaking compliance language may be relevant but not sufficient.

Another finding concerns the potential risk of concealment. Third-party exposure of undisclosed AI use can damage trust even more than voluntary disclosure. This means concealment can seem attractive in the short run, but it could become a high-risk gamble should AI use become easier to detect.[1]

A second set of experiments examined why disclosure lowers trust. In these studies, participants evaluated AI-assisted work and then judged whether they trusted the actor and whether the AI use seemed legitimate for the setting. The results pointed to legitimacy—the perception that an action is appropriate for its social setting[8,9]—as the key pathway: Specifically, disclosure heightened concerns about typicality, commitment, and authenticity. That is, once AI use was disclosed, audiences asked whether AI use is normal in this setting (typicality), whether the person put in the required effort (commitment), and whether a human still genuinely owned the output (authenticity). Otherwise identical work came to seem less legitimate on all three counts, and these reactions helped to explain the drop in trust.[1,2] Table 1 summarizes key empirical findings from our research, alongside legitimacy concerns and their implications for policy.

Disclosure penalties are not unique to AI. The broader literature on disclosure shows that transparency can have unintended consequences to the extent that disclosures function as negative cues about true motives or responsibility. Conflict of interest disclosures, for example, can reduce trust in advice, even when disclosure is intended to protect decision makers.[11,12]

Policy Recommendations: A Trust-Preserving AI Disclosure Standard

Our article addresses a broad design problem for regulators and employers: how to establish legitimate AI-assisted practices while preserving transparency. The challenge cuts across many domains: workplace and HR communications,

Table 2. The trust-preserving AI disclosure standard at a glance.

Recommendation	What it looks like in practice
Normalize approved AI use first	Publish permitted uses and approved tools and have leaders visibly use them in everyday work.
Tier disclosure obligations according to the stakes	Keep it brief for routine uses, add a short scope note for consequential ones, and require a documented decision trail for high-stakes decisions.
Pair disclosure with ownership and accountability	Attach a short note stating what the AI did, what the human did, and who owns the result.
Enforce consistently	Apply mandates predictably and protect good-faith disclosure so honesty is not penalized.

education, consumer and marketing messaging, health care, and other high-stakes organizational decisions. The proposed standard reorients the policy focus from the binary question of whether to disclose AI use toward designing disclosures that preserve transparency while minimizing unnecessary legitimacy costs.[1,2,7] Table 2 summarizes key aspects of the proposed disclosure standard as a concise checklist for organizations and policymakers.

Establish Norms for Acceptable AI Use First

Organizations should establish norms for acceptable AI use before (or at least alongside) disclosure mandates. Leaders and regulators can help create legitimacy by clarifying acceptable uses of AI, releasing lists of approved tools, and describing compliant AI use in routine workflows. When AI assistance is framed as a standard and role-consistent practice rather than as a secretive shortcut, disclosure is less likely to be interpreted as improper.[1,2,10,13] However, these norms might take time to spread across people and organizations and might not show immediate results.

Tier Disclosure Obligations According to the Stakes

Disclosure obligations should differ according to the stakes involved in a given task. When AI assistance is routine or technical and concerns low-stakes decisions, simple disclosure might suffice. Examples may include using AI to edit grammar, create internal summaries, or schedule meetings. In contrast, full disclosure about human ownership might be necessary for high-stakes decisions that concern opportunities, rights, or personalized judgment. Such disclosure should include a review process and appeal language. We suggest a three-tier framework tied to stakeholder risk. A tiered approach keeps routine disclosure simple while requiring more details for consequential uses.[6,7]

- Tier 1 (routine and low-stakes decisions): AI used for drafting, summarizing, or editing where errors have low consequences (e.g., internal scheduling messages). Use a brief standardized label plus a short ownership and accountability addendum (see "Pair the Disclosure With a Brief Statement of Ownership & Accountability.")
- Tier 2 (consequential communications): AI used in communications that can materially shape people's beliefs or evaluations (e.g., client recommendations, graded feedback, or performance evaluations).[14] Add a short note stating what AI did and did not do (e.g., "AI was not used to generate data or final decisions").
- Tier 3 (high-stakes decisions): AI used in decisions affecting rights, opportunity, safety, or access (e.g., hiring, promotion, termination, pay, credit, insurance, housing, AI-assisted diagnosis or triage, or access to health care). Require Tier 2 disclosure plus a documented history of how decisions were made, including the tool used, the purpose, who the human reviewer of record was, and a summary of the decision-making process sufficient for later review and dispute resolution.

To give some concrete examples, a brief label may be sufficient for Tier 1 uses: "AI assisted with drafting or editing this message. The final text was reviewed and approved by [name/role]." For Tier 2 uses, the disclosure should also identify scope: "AI helped generate initial language or options. It did not generate the underlying data, evaluation, or final judgment. [Name/role] reviewed the relevant information, revised the content, and is accountable for the final message." For Tier 3 uses, the disclosure should also preserve a review record: "AI was used for [purpose] with [tool/system]. [Name/role] reviewed the output, made the final decision, and can be contacted through [appeal or escalation path]. A record of the decision process is available for appropriate review."

In summary, tiering should support AI adoption without treating transparency as a compliance burden that discourages legitimate use. It keeps disclosure practical while protecting stakeholders where the stakes are high.[15]

Pair the Disclosure With a Brief Statement of Ownership & Accountability

To counter inferences about low commitment and low authenticity, Tier 2 and 3 disclosures should include a short,

structured addendum that clearly identifies ownership and accountability[2]:

- What the AI did (e.g., the AI drafted, summarized, and/or generated options).
- What the human did (e.g., the person set goals, provided inputs, checked key claims, edited, and/or made the final decision).
- Who is accountable (e.g., the person's name and role; for Tier 3 decisions, include an appeal path).

This advances disclosure from a binary label ("AI was used") to an accountability statement ("Here is what happened and who owns it"). This added clarity should reduce ambiguity about authenticity, a key legitimacy fault line that can drive distrust.

Make Enforcement & Compliance Predictable

Where feasible, regulators should back disclosure requirements with enforcement (especially for Tier 3) to prevent adverse selection. Employers should provide standardized templates and, where possible, include default disclosure footers in common tools so disclosure feels routine rather than stigmatizing. Good-faith disclosure of permitted AI use should not itself trigger discipline. Focus enforcement on deceptive nondisclosure, known misuse of unapproved tools, and failure to complete required human review steps. [1,2,6,7]

Application of the Policy Recommendations Across Domains

Effective implementation of transparency standards will differ across contexts. The appropriate disclosure tier and accompanying safeguards should reflect the role AI plays in the task and the potential consequences for affected stakeholders. The following examples illustrate how the proposed standard can be adapted to different organizational and professional settings.

- Workplace and management: The four steps should be integrated into ordinary AI governance rather than treated as an exception-handling requirement. Approved AI uses and auditing responsibilities should be defined by senior leaders. All-staff sessions can help set expectations, share templates, and discuss reporting mechanisms. Line managers can use the Tier 2 disclosure and the ownership and accountability addendum for coaching or performance reviews. Tier 3 disclosure should apply to hiring, promotion, termination, pay, and other employment decisions. Employers may choose to require that certain communications, such as termination letters or tailored coaching, be written entirely by humans or involve disclosure that is more extensive than Tier 2 standards.
- Education: Tier 1 should cover editing and translation assistance. Tier 2 should cover feedback or evaluation-related communications, and Tier 3 should cover admissions, credentialing, and integrity decisions. Statements of ownership should clarify what is assistance and what is authorship.
- Research: Copyediting, translation, and formatting fall under Tier 1. AI-assisted drafting, literature synthesis, and code assistance fall under Tier 2, with a scope note (see the AI disclosure for this article as an example). Data analysis bearing on findings, peer review, and integrity decisions require Tier 3. Ownership statements should distinguish assistance from authorship, with only humans qualifying as authors.
- Marketplace: Persuasive messaging or personalization driven by AI should typically employ Tier 2 disclosure, positioned close to any claim it makes, accompanied by a scope note. Financial and legal advice-giving services should use Tier 3 disclosure when AI materially informs personalized advice or recommendations that may affect individuals' rights, opportunities, or significant financial or legal interests.
- Health care: For patient-facing AI uses in imaging interpretation, triage, or portal messaging, disclosures should identify the tool's role and clinician review, not merely state that AI was used. Routine communications may fit Tier 2; AI tied to diagnosis, treatment, or access to care should use Tier 3.

What We Still Do Not Know

Most research to date has examined swift trust; that is, the trust people extend in initial encounters before any interaction history has accumulated.[16] Because audiences meeting an actor for the first time cannot draw on firsthand experience, swift trust rests on presumptive cues such as the actor's role and reputation, institutional safeguards, and the expectation that the actor will behave in ways typical of the setting.[16] This is what makes first impressions of AI use so consequential. In an initial encounter, an AI disclosure is one of the few concrete cues available; absent established norms, it can read as a deviation from expected practice, undermining exactly the category-based assumptions on which swift trust is built. Our studies, which involved unfamiliar actors in controlled settings, largely captured this early stage of trust formation. Trust penalties may be smaller in long-standing relationships where trust is anchored in an accumulated track record rather than in single cues. However, organizations should not assume that familiarity always eliminates the problem. This set of recommendations should be seen in context as the evidence base in an AI-mediated world grows.

Conclusion

Mandating transparency about AI use is ethically appealing. However, recent empirical evidence shows that transparency can systematically erode trust by triggering legitimacy concerns.[1,2] The recommendations outlined herein can preserve the benefits of transparency while also protecting the trust that enables organizations to function effectively.

ORCID iD

Martin Reimann https://orcid.org/0000-0002-3246-1878

Funding

Research support was provided by a National Science Foundation CAREER Award (No. 1943688) to the first author and by a National Endowment for the Arts research grant (No. 1925643-38-24) and an AI and Health Opportunity Seed Grant from the University of Arizona to the second author. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Author AI Disclosure Note (Tier 2)

AI assisted: Generative AI writing tools were used to assist with drafting, revising, and copyediting portions of this manuscript for grammar, clarity, organization, and style. Scope: AI tools were not used to generate the manuscript's ideas, arguments, conclusions, or policy recommendations. Human ownership and accountability: The coauthors developed the arguments and recommendations, directed the AI-assisted writing process, evaluated and selectively adopted AI-suggested text, independently checked factual claims and citations, made all final decisions about the content, and are accountable for the manuscript and any remaining errors.

References

1. Schilke O, Reimann M. The transparency dilemma: how AI disclosure erodes trust. *Organ Behav Hum Decis Process*. 2025;188:104405. doi:10.1016/j.obhdp.2025.104405
2. Schilke O, Reimann M. How does AI disclosure shape trust? Unpacking the role of legitimacy. *Soc Psychol Q*. Published online April 14, 2026. doi:10.1177/01902725261425997
3. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. *Acad Manage Rev*. 1995;20(3):709-734. doi:10.5465/amr.1995.9508080335
4. Dirks KT, de Jong B. Trust within the workplace: a review of two waves of research and a glimpse of the third. *Annu Rev Organ Psychol Organ Behav*. 2022;9(1):247-276. doi:10.1146/annurev-orgpsych-012420-083025
5. Schilke O, Reimann M, Cook KS. Trust in social relations. *Annu Rev Sociol*. 2021;47:239-259. doi:10.1146/annurev-soc-082120-082850
6. Schnackenberg AK, Tomlinson EC. Organizational transparency: a new perspective on managing trust in organization-stakeholder relationships. *J Manage*. 2016;42(7):1784-1810. doi:10.1177/0149206314525202
7. El Ali A, Venkatraj KP, Morosoli S, Naudts L, Helberger N, Cesar P. Transparent AI disclosure obligations: who, what, when, where, why, how. In: *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery; 2024:Article 342. doi:10.1145/3613905.3650750
8. Suchman MC. Managing legitimacy: strategic and institutional approaches. *Acad Manage Rev*. 1995;20(3):571-610. doi:10.5465/amr.1995.9508080331
9. Bitektine A, Haack P. The "macro" and the "micro" of legitimacy: toward a multilevel theory of the legitimacy process. *Acad Manage Rev*. 2015;40(1):49-75. doi:10.5465/amr.2013.0318
10. Haack P, Schilke O, Zucker LG. Legitimacy revisited: disentangling propriety, validity, and consensus. *J Manag Stud*. 2021;58(3):749-781. doi:10.1111/joms.12615
11. Loewenstein G, Cain DM, Sah S. The limits of transparency: pitfalls and potential of disclosing conflicts of interest. *Am Econ Rev*. 2011;101(3):423-428. doi:10.1257/aer.101.3.423
12. Sah S, Malaviya P, Thompson D. Conflict of interest disclosure as an expertise cue: differential effects due to automatic versus deliberative processing. *Organ Behav Hum Decis Process*. 2018;147:127-146. doi:10.1016/j.obhdp.2018.05.008
13. Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. *Manag Sci*. 2018;64(3):1155-1170. doi:10.1287/mnsc.2016.2643
14. Reimann M, Kronrod A. Language as means and ends: how generative artificial intelligence automates, amplifies, and reinvents language in marketing. *J Assoc Consum Res*. 2026;11(2):227-234. doi:10.1086/740067
15. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. *Acad Manag Ann*. 2020;14(2):627-660. doi:10.5465/annals.2018.0057
16. Meyerson D, Weick KE, Kramer RM. Swift trust and temporary groups. In: Kramer RM, Tyler TR, eds. *Trust in Organizations: Frontiers of Theory and Research*. Sage; 1996:166-195. doi:10.4135/9781452243610.n9